Cascabel: a flexible, scalable and easy-to-use amplicon sequence data analysis pipeline

Alejandro Abdala Asbun[1,✉], **Marc A Besseling**[1], **Sergio Balzano**[1,2], **Judith van Bleijswijk**[1], **Harry Witte**[1], **Laura Villanueva**[1], **and Julia C Engelmann**[1,✉]

[1]Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research and Utrecht University, P.O. Box 59, 1790 AB, Den Burg, The Netherlands
[2]Current address: Stazione Zoologica Anton Dohrn, Instituto Nazionale di Biologia, Ecologia e Biotecnologie Marine, Villa Comunale, 80121, Naples, Italy

ABSTRACT

Marker gene sequencing of the rRNA operon (16S, 18S, ITS) or cytochrome c oxidase I (CO1) is a popular means to assess microbial communities of the environment, microbiomes associated with plants and animals, as well as communities of multicellular organisms via environmental DNA sequencing. Since this technique is based on sequencing a single gene rather than the entire genome, the number of reads needed per sample is lower than that required for metagenome sequencing, making marker gene sequencing affordable to nearly any laboratory. Despite the relative ease and cost-efficiency of data generation, analyzing the resulting sequence data requires computational skills that may go beyond the standard repertoire of a current molecular biologist/ecologist. We have developed *Cascabel*, a flexible and easy-to-use amplicon sequence data analysis pipeline, which uses Snakemake and a combination of existing and newly developed solutions for its computational steps. *Cascabel* takes the raw data as input and delivers a table of operational taxonomic units (OTUs) and a representative sequence tree. Our pipeline allows customizing the analyses by offering several choices for most of the steps, for example different OTU generating methods. The pipeline can make use of multiple computing nodes and scales from personal computers to computing servers. The analyses and results are fully reproducible and documented in an HTML and optional pdf report. *Cascabel* is freely available at Github: https://github.com/AlejandroAb/CASCABEL and licensed under GNU GPLv3.

amplicon sequencing, 16S/18S rRNA gene, Illumina, Snakemake, community profiling, microbiome, pipeline

Correspondence: `alejandro.abdala@nioz.nl`, `julia.engelmann@nioz.nl`

Introduction

High-throughput sequencing of an omnipresent marker gene such as the gene coding for the small subunit of the ribosomal RNA (16S for prokaryotes or 18S for eukaryotes) is a cheap means for microbial community profiling that is affordable for nearly every lab. Moreover, sequencing a marker gene like cytochrome c oxidase I (CO1) in environmental DNA also allows to track larger multicellular organisms, for example fish in the sea. Amplicon sequencing can also be used to investigate active microbial communities based on ribosomal RNA abundance instead of the rRNA gene locus (1, 2). Typically, a short fragment of 100-600 nucleotides of the marker gene with the desired taxonomic resolution is amplified by PCR from the DNA extract or cDNA of the community and then sequenced by high throughput sequencing. On current sequencing platforms, up to hundreds of samples can be combined (multiplexed) in a single sequencing run, decreasing the sequencing costs per sample tremendously. Not surprisingly, community compositions based on DNA analyses have been generated from most of the habitats on earth, including the human body, e.g., (3), the open ocean, e.g., (4), deep sea, e.g., (5) and intracellular symbionts, e.g., (6). The current bottleneck in studies using community profiling is the computational analysis of the (potentially massive) sequence data. For scientists with little background in bioinformatics, the amount of data and complexity of data analysis can be overwhelming. Several software solutions for the individual steps from raw sequence data to an operational taxonomic unit (OTU) table exist (7, 8), but are not necessarily straightforward to use. The software package Mothur (7) which comes with its own computational environment and the QIIME framework (8) have been popular platforms for data analysis, but both solutions require the ability to work on the command line. Analyzing multiple sequencing libraries quickly becomes tedious for users not proficient in implementing bash (or any other programming language) scripts which chain the individual steps and allow parallel processing. While web servers for microbial community data analysis like SILVAngs (9) and MG-RAST (10) are easy to use, they are inherently inflexible and also limited in throughput. QIIME2 (11) has command line and graphical user interface modes of operation and offers even a larger choice of algorithms for data analysis than the original QIIME, including statistical analyses of the resulting community profiles. We anticipated a need for a pipeline which combines the flexibility provided by using bioinformatic tools on the command line within one of the existing frameworks with the ease of using interactive web servers for analyzing and interpreting amplicon sequencing data.

We here provide *Cascabel*, a Snakemake (12) pipeline for the analysis of community marker gene sequence data which is easy to use for people with little bioinformatics background, and both flexible and powerful enough to be attractive for people with bioinformatics training. *Cascabel* supports large sample and sequencing library throughput as well as parallel computing on personal computers and computing servers. Moreover, all input and output files, tools, parameters and their versions are documented and render the analyses fully reproducible.

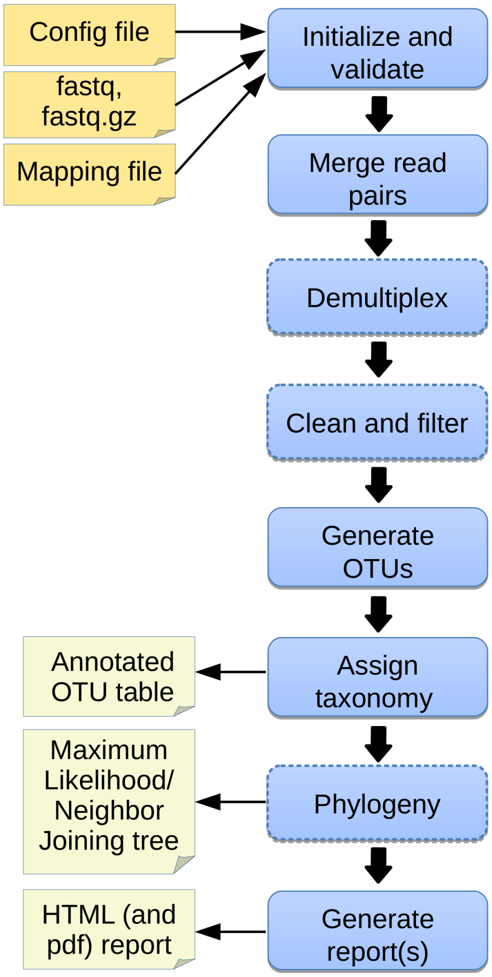

Fig. 1. Overview of *Cascabel*. The workflow indicates input files (config file, sequence data in fastq format, barcode mapping file), mandatory and optional steps of the pipeline (blue boxes) as well as the main output files. The boxes of optional steps have dashed borders. 'Clean and filter' refers to removing primers/adapters and chimeras. Table 1 shows a detailed summary of the steps, available tools and output files.

Implementation

Our pipeline makes use of the workflow management engine Snakemake (12), which scales from personal workstations to compute clusters. *Cascabel* consists of a set of 'rules', which specify the input, the action to perform on the input (executed by a bash/python/R/java script), and the output. The user defines via a configuration file (called 'config file' from now on) in yaml format, how these 'rules' are chained to perform amplicon sequence data analysis from the raw data to the final OTU table. For most of the rules *Cascabel* provides several alternative algorithms or tools and allow passing arguments via the config file to the algorithm being used. In addition, rules can be skipped, and the pipeline can be entered and exited at every step. This makes *Cascabel* very flexible and highly customizable. Moreover, the pipeline is easily extendable and amendable to personal needs, allowing e.g. the analysis of any marker gene sequence data. Running *Cascabel* requires the raw fastq sequence data files, a mapping file indicating which sample carries which barcode if the data should be demultiplexed, and optionally sample metadata (e.g., geographic coordinates of the sampling stations, physical, chemical, or biological properties), and the config file specifying the tools and parameters used for running the pipeline. We provide default parameters but strongly advise making informed choices about parameter settings matching the individual needs of the experiment and data set. With the files in place, *Cascabel* is started with a one-line command on the terminal. Snakemake takes care of executing the rules in a computationally efficient manner, making optimal use of available resources e.g., distributing jobs over several nodes. Figure 1 provides an overview of the workflow of *Cascabel*.

Cascabel supports paired-end sequence data as input from one or multiple samples per input file. Barcodes for demultiplexing samples can be situated at the beginning of one or both of the reads.

Cascabel provides a range of popular methods to generate OTUs with or without a reference sequence database (swarm (13), sortmerna (14), mothur (7), trie (QIIME team, unpublished), uclust/uclust_ref/usearch/usearch_ref (15), prefix/suffix (QIIME team, unpublished), cd-hit (16), sumaclust (17)) which are executed by QIIME. Then, representative sequences are chosen for each OTU (with options: random, longest, most_abundant, first) (8).

Table 1. Outline of the steps performed by *Cascabel*. 'script(s)' refers to *Cascabel* scripts in bash, java or R.

Step	Tools/Algorithms	Output
Initialize structure	script	Project folder and file structure
Quality Control	FastQC(18)	FastQC report
Merge reads	PEAR(19)	Merged (assembled) sequences
Demultiplex	QIIME(8), scripts	Sequences assigned to samples in one file and per sample
Align versus reference	mothur(7)	Aligned sequences
Remove chimeras	usearch61(15)	Chimera-free sequences
Remove adapters	Cutadapt(20)	Adapter-free sequences
Size filter	script	Filtered sequences
Dereplicate	VSEARCH(21)	Dereplicated sequences
Generate OTUs	VSEARCH, swarm(13), uclust(15), trie, sortmerna(14)	OTU table
Assign taxonomy	QIIME (BLAST(22), uclust, RDP(23)), blastn (BLAST+)(24), VSEARCH	Taxonomic assignments for each OTU
Generate OTU table	QIIME	Annotated OTU table
Alignment	pynast(25), mafft(26), infernal(27), clustalw(28), muscle(29)	multiple sequence alignment
Make tree	muscle, clustalw, raxml(30), fasttree(31)	phylogenetic tree
Report	scripts, Krona(32)	HTML, pdf report, Krona charts

Cascabel assigns taxonomy to the representative sequences using the in-built most recent version of the SILVA database, at this moment v132 (9), providing three different approaches: VSEARCH, which performs global alignment of the target sequences against the reference database; BLAST, making use of BLAST+ (24); QIIME, with methods BLAST (22), uclust or the RDP classifier. Alternatively, any other public or custom database can be used for taxonomic annotation. If taxonomy is assigned with VSEARCH or BLAST, the user can choose to assign the sequences to the lowest common ancestor (LCA) with the stampa approach (https://github.com/frederic-mahe/stampa). The last rule of *Cascabel* (the 'target' rule) generates HTML and optional pdf reports with documentation, figures and tables summarizing the results of individual rules, as well as all software versions used. Table 1 shows an overview of the options and methods provided for the individual steps of the analysis performed by *Cascabel*.

Cascabel has an interactive and a non-interactive mode. In interactive mode, several modules have a check-point which needs to be passed to continue with the analysis. If the check fails (e.g., if too many FASTQC quality modules failed or the number of sequences assigned to sample barcodes is too low), the pipeline stops and the user has to either change parameters and continue, or exit the pipeline. If parameters were changed interactively, the new ones are documented in the reports.

Snakemake will not re-run a rule if the output file of that rule already exists. This avoids over-writing existing results, but also renders it impossible to keep results of multiple analyses on the same data in the same project. To avoid initializing multiple projects with the same raw data, we implemented *Cascabel* with a 'Run' parameter. Whenever the user changes the Run parameter, a new analysis will be performed (except for quality control on the raw data) and the results saved in a different 'Run' folder. Moreover, the user can perform taxonomic assignments for the same run using different methods and the results will be saved in individual 'taxonomy' folders. When starting a new taxonomic assignment, the existing OTU representative sequences are used so no processing time is wasted by performing the same upstream rules several times.

Results

To show the utility of our pipeline, we applied it to 16S rRNA gene amplicon data generated from water column samples taken from Lake Chala, which is situated on the border of Kenya and Tanzania, east of Mount Kilimanjaro in Africa. As stated earlier, however, the pipeline can process sequence data from any marker gene. *Cascabel* comes with taxonomic mapping files for 16S rRNA and 18S rRNA gene sequences from SILVA v132, but the user can always choose to make use of a different public or a custom reference sequence database.

First, *Cascabel* checked the validity of the input files including the barcode mapping file and the config file. Supplementary file 1 contains the config file of the analyses presented here. After having validated the input files, *Cascabel* proceeded with analyzing sequence data quality with FastQC (18). In interactive mode, *Cascabel* will stop if more than a specified number of quality check modules failed. Next, we assembled read pairs using PEAR (19) and assessed the quality of the assembled reads with FastQC.

When working with large datasets, a dereplication step which collapses identical sequences into one representative sequence can drastically reduce computation time. We provide a custom rule based on VSEARCH (21) which keeps track of the abundance of the individual sequence across samples. We dereplicated sequences which were identical over the full sequence length. If the library contains sequences from several samples, they are then demultiplexed based on the barcode sequences provided in the barcode mapping file. To do so, we make use of QIIME (8) and a custom R script to (optionally) allow sequence errors in the barcodes. Demultiplexed data can also be stored in individual fastq files for further use outside the pipeline, e.g., for submitting data to public repositories. We demultiplexed the example 16S rRNA gene sequencing data based on a sample barcode of 12 nucleotides located at the beginning of the forward read. Optionally, *Cascabel* will align sequence reads against a reference sequence database to facilitate removing sequence adapters or primers or both. Adapter and primer sequences can be trimmed off with Cutadapt (20). We did not apply Cutadapt on our example data set because we did not expect adapters and did not consider primer removal necessary for the purpose of demonstrating the features of *Cascabel*. Then, *Cascabel* generates a histogram of sequence lengths. In interactive mode, *Cascabel* shows the frequency of occurrence of each of the read lengths on the terminal and allows to change the minimum and maximum sequence length provided in the config file. For our 16S data set, we filtered out sequences whose length differed by more than 15 nucleotides from the average sequence length. The library report contains a smoothed histogram of the sequence lengths to validate the choice of the minimum and maximum sequence length (Figure 2A). Optionally, *Cascabel* identifies and removes chimeras either *de novo* based on sequence abundance or searching against the gold database provided by QIIME with the usearch61 algorithm (15). The user can also provide a different databases such as SILVA (9) or PRD2 (33) to search chimeras. Assembled and potentially filtered sequence reads from all samples are then concatenated into one fasta file. *Cascabel* then generates a histogram to visualize the number of reads per sample for the report of the complete run including all libraries (named 'otu_report') to assess whether the sequences are evenly spread across the samples (Figure 2B). In the data set analyzed, indeed, the number of reads per sample vary to some extent. This is observed frequently even though amplicons from the different samples were pooled in equimolar amounts. Furthermore, the reports for each of the libraries contain a plot of the number and percentages of raw, assembled, demultiplexed and length filtered sequences (Figure 2C).

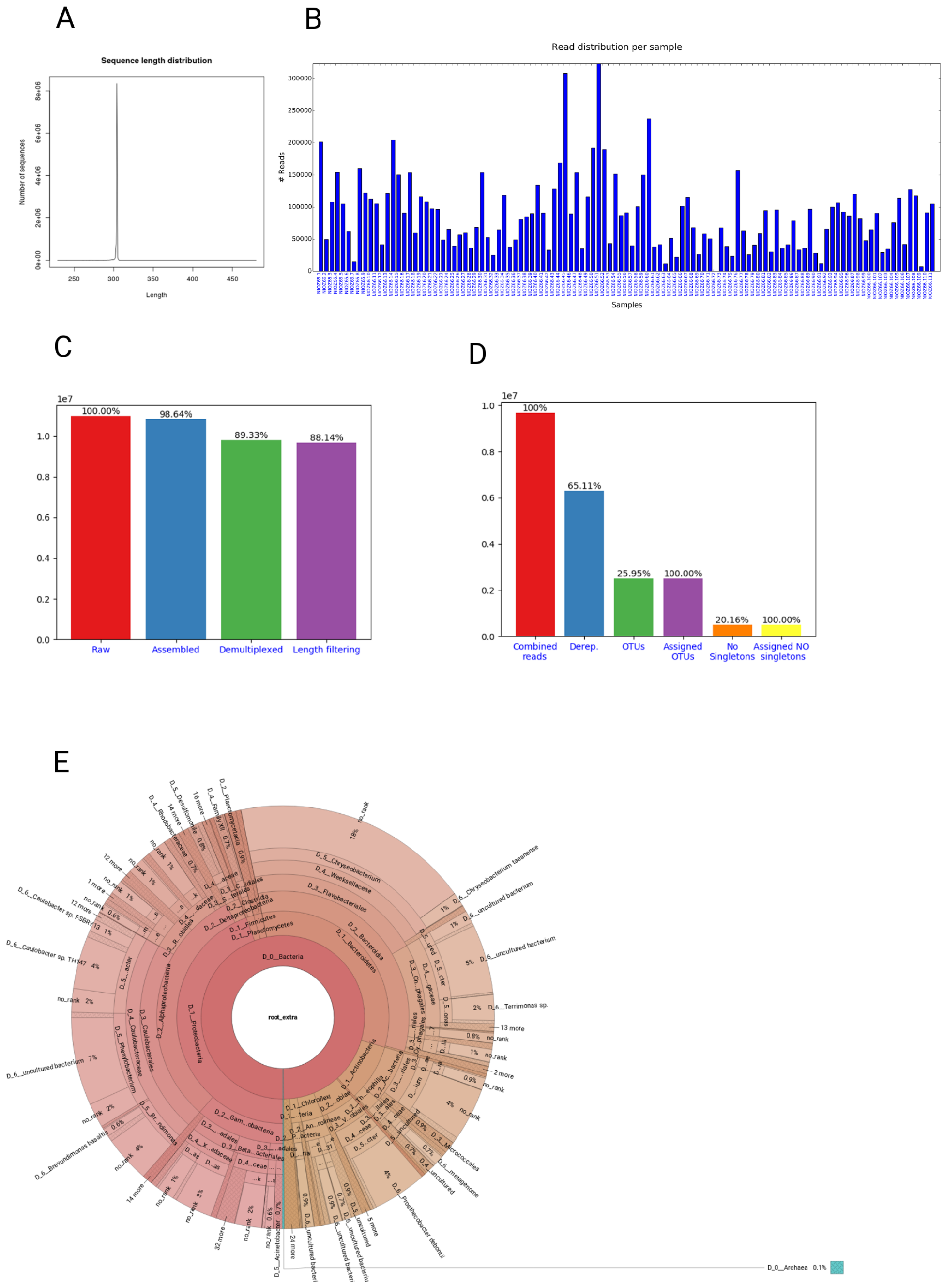

Fig. 2. Figures shown in *Cascabel* reports. **(A)** Smoothed sequence length distribution after merging reads, for one library. The plot is meant to help making a sensible choice for sequence length filtering. **(B)** Number of sequences per sample. This histogram is part of the OTU report (including all libraries). **(C)** Number of sequences after individual preprocessing steps. 'Assembled' refers to the number of read pairs which could be merged based on their overlap. This plot is part of the library report. **(D)** Number of sequences after individual steps after potentially combining several libraries (total number of reads) and generating OTUs. 'Derep.' indicates the number of dereplicated reads. 'OTUs' is the total number of OTUs and 'Assigned OTUs' the number of OTUs with a taxonomic assignment. 'No singletons' refers to the number of OTUs excluding singleton OTUs and 'Assigned NO singletons' the number of singleton-free OTUs with a taxonomic assignation. The plot is part of the OTU report. **(E)** Krona chart for one sample. The krona charts are interactive and can be viewed with a web browser. Colors indicate the taxonomic groups that the OTU was assigned to. Each ring of the pie chart represents a different taxonomic level. An example of a full library report is shown in supplemental file 2, and an OTU report is provided in supplemental file 3.

Next, *Cascabel* will cluster reads from the sequence data of all libraries into OTUs. We used 97% sequence identity with uclust to generate roughly 2.5 million OTUs. We chose the longest sequence of an OTU as representative sequence to be used for taxonomic placement of the OTU. Alternative OTU and representative sequence picking methods provided by *Cascabel* are listed in Table 1. Then we used VSEARCH to assign taxonomy to the representative sequences based on the SILVA database (SILVA version 132).

From the abundances of the OTU sequences within each of the samples, *Cascabel* creates an OTU abundance table. The OTUs can further be grouped at higher taxonomic levels depending on the desired resolution. Subsequently, the user can opt to remove singletons, align representative sequences, filter the alignment and make a phylogenetic tree. Removing singletons reduced the number of OTUs in the analyzed dataset to roughly 500.000. To align representative sequences, *Cascabel* offers pynast (25), mafft (26), infernal (27), clustalw (28), and muscle (29), and we used pynast on our data. A phylogenetic tree can be generated with muscle, clustalw, raxml (30) and fasttree (31), and we applied fasttree (Table 1).

Finally, *Cascabel* generates HTML and optionally pdf reports of the analyses, documenting all software and parameters used. If more than one library was analyzed, there will be a report for each library as well as a report summarizing all libraries (otu_report). Among other graphics, the otu_report shows the percentages and the total number of reads after filtering ('combined reads'), dereplicated reads, OTUs, OTUs assigned to a taxonomic level, OTUs excluding singletons ('no singletons'), and assigned OTUs excluding singletons. The graph for the analyzed example data is shown in Figure 2D. Supplementary files 2 and 3 show the library and the otu_report for the demonstration data, respectively. In addition, *Cascabel* generates an interactive Krona chart (32) for the run which displays community composition for individual samples or the complete data set. The Krona chart shows the taxonomic assignments in an interactive HTML document composed of a multi-layered pie-chart and the user can zoom and browse these different levels. An example is shown in Figure 2E.

The user can make use of all intermediate files generated by individual rules, and most importantly the OTU table and representative sequences for follow-up analyses. To save disk space, the user can also opt to have *Cascabel* remove temporary files at the end of the analyses. For many rules, the user can pass additional parameters to the command or tool at hand using the 'extra_params' parameter in the config file.

Discussion

Cascabel has been developed at the Royal Netherlands Institute for Sea Research (NIOZ) to facilitate, unify and easily track data provenance of amplicon sequence data analyses. Apparent advantages of using this pipeline compared to custom scripts are that the individual steps of the pipeline have been tested by many members of the community at the NIOZ who are experienced in amplicon sequencing data analyses (34–37), and therefore should contain fewer mistakes than scripts that were written for a specific analysis by one person. Moreover, community knowledge and experience have created a workflow which is probably more comprehensive and powerful than one that was created by a single person. In addition, the availability of the pipeline has facilitated comparing and integrating research results from different data sets generated at the NIOZ because scientists can agree on certain settings and reference database versions and the pipeline guarantees that the analyses are performed in the same way. Because *Cascabel* keeps track of data provenance, documenting the process of analyzing the data to generate results, it also facilitates preparing research manuscripts. While most of the scientific journals request the raw sequencing data to be submitted to a public repository for many years already, also reporting data provenance becomes more important. The journal 'Nature', for example, requires authors to make materials, data, code, and associated protocols available (38). *Cascabel* facilitates providing data, code and protocols. Public sequence repositories often require the raw data to be submitted per sample, but sample demultiplexing typically takes place after merging read pairs such that the raw data cannot be recovered. Therefore *Cascabel* demultiplexes the raw data in parallel to the analyses such that it is ready for public data repository deposition. The code of Cascabel is open source and all analyses are protocoled in the reports and config file, complying with the rules for reproducible computational research described by Sandve et al. (39).

DNA sequencing technology, algorithms and analysis approaches are constantly evolving. It is logical that pipelines lag behind with the most recent developments because it takes time to test and integrate new modules. Because *Cascabel* is a Snakemake workflow, it is flexible and easy to extend to encompass more or alternative rules. We are constantly working on extending the range of applications and making alternative approaches, like generating amplicon sequence variants (ASVs) instead of OTUs, available. *Cascabel* provides reference databases for taxonomy assignment and chimera detection, but the user can always supply a different database and specify that in the config file. Moreover, *Cascabel* is not limited to Illumina sequence data that we used for demonstration purposes, but can handle sequence data from other technologies which produce short reads from amplicons as well (e.g. Ion Torrent). With some minor modifications, *Cascabel* can even be used to analyze long read amplicon sequence data.

Galaxy (40) might be a user-friendly web-based alternative to *Cascabel* which offers interfaces to VSEARCH and mothur executables. Having a medium-sized user group at the institute, we did not want to overload a public server and setting up and maintaining our own server would also need resources that we preferred to allocate to the development of a workflow for which we have full control and flexibility. With *Cascabel* being invoked from the command line, the user can make use of the full potential that Snakemake has to offer, e.g., `-prioritize` to force the execution of specific rules prior to others when distributing tasks across computing resources, `-until` to run the pipeline up to a specific rule, `-summary`, which shows the rules executed so far and `-dag` which shows the rules executed and the ones yet to be done in a directed acyclic graph. Moreover, we consider *Cascabel*'s report an essential element to move forward in terms of user-friendly data provenance and reproducibility.

We have presented *Cascabel*, an open source pipeline to analyze amplicon sequence data based on the workflow engine Snakemake. The pipeline can be easily installed via conda, comes with documentation and a wiki on github and can be executed by users with basic command line skills. At the same time, *Cascabel* is flexible, offering alternative options for most of the steps and supporting custom reference databases, and can easily be modified and extended by users with computational skills. We believe that *Cascabel* will prove to be useful to scientist who need more flexibility and throughput than provided by tools based on web servers, but do not want to or cannot generate their own command-line based workflow.

Methods

Sampling and DNA extraction. Suspended particulate matter (SPM) was collected from Lake Chala from September 2013 to May 2014 from a total of 111 samples as described by (41). DNA was extracted from 1/32 section of the filters on which SPM was collected by using the PowerSoil DNA extraction kit (Mo Bio Laboratories, Carlsbad, CA, USA).

DNA sequencing. The V4 region of the 16S rRNA gene were amplified with the primers forward:
515F (Parada): GTGYCAGCMGCCGCGGTAA (42) and reverse:
806R (Apprill): GGACTACNVGGGTWTCTAAT (43). We made use of 12 nucleotide Golay barcodes at the beginning of the forward read. Paired-end sequencing of 250 nt was performed on an Illumina MiSeq instrument (Illumina, San Diego, CA) using the Truseq DNA nano LT kit for library preparation and V3 sequencing chemistry at the sequencing facility of the University of Utrecht (USEQ), the Netherlands. The data is publically available at NCBI, BioProject PRJNA526242.

Sequence analysis. All the settings and parameters chosen to analyze the example data set are given in the config file (supplementary file 1) and the reports (supplementary files 2 and 3).

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

AA implemented the pipeline, with contributions from JCE. MB, LV, SB and HW tested the pipeline. AA, JvB and JCE designed the pipeline, with contributions from LV. JCE wrote the manuscript. All authors contributed to and approved the final version of the manuscript.

Funding

AA was supported by the Soehngen Institute of Anaerobic Microbiology (SIAM) Gravitation grant (024.002.002) of the Netherlands Ministry of Education, Culture and Science (OCW) and the Netherlands Organisation for Scientific Research (NWO).

Acknowledgments

We thank Hans Malschaert for IT support and the bioinformatics community at the NIOZ for comments and suggestions about the pipeline. We acknowledge S. Vreugdenhil and M. Brouwer for technical and analytical support. Fieldwork with collection of the studied sample materials was carried out with permission from the government of Kenya through permit 13/001/11C to D.V. In accordance with National Environmental Management Authority regulations in the context of the Nagoya Protocol, DNA extracts of the analyzed suspended-particulate samples are archived at the National Museums of Kenya (NMK), under voucher numbers NMK:BCT:80001 to NMK:BCT:80221.

Bibliography

1. Ramon Massana, Angélique Gobet, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Aurélie Chambouvet, Richard Christen, Jean-Michel Claverie, Johan Decelle, John R Dolan, Micah Dunthorn, Bente Edvardsen, Irene Forn, Dominik Forster, Laure Guillou, Olivier Jaillon, Wiebe H C F Kooistra, Ramiro Logares, Frédéric Mahé, Fabrice Not, Hiroyuki Ogata, Jan Pawlowski, Massimo C Pernice, Ian Probert, Sarah Romac, Thomas Richards, Sébastien Santini, Kamran Shalchian-Tabrizi, Raffaele Siano, Nathalie Simon, Thorsten Stoeck, Daniel Vaulot, Adriana Zingone, and Colomban de Vargas. Marine protist diversity in european coastal waters and sediments as revealed by high-throughput sequencing. *Environmental microbiology*, 17:4035–4049, October 2015. ISSN 1462-2920. doi: 10.1111/1462-2920.12955.
2. Dominik Forster, Micah Dunthorn, Fréderic Mahé, John R Dolan, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Richard Christen, Jean-Michel Claverie, Johan Decelle, Bente Edvardsen, Elianne Egge, Wenche Eikrem, Angélique Gobet, Wiebe H C F Kooistra, Ramiro Logares, Ramon Massana, Marina Montresor, Fabrice Not, Hiroyuki Ogata, Jan Pawlowski, Massimo C Pernice, Sarah Romac, Kamran Shalchian-Tabrizi, Nathalie Simon, Thomas A Richards, Sébastien Santini, Diana Sarno, Raffaele Siano, Daniel Vaulot, Patrick Wincker, Adriana Zingone, Colomban de Vargas, and Thorsten Stoeck. Benthic protists: the under-charted majority. *FEMS microbiology ecology*, 92, August 2016. ISSN 1574-6941. doi: 10.1093/femsec/fiw120.
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486:207–214, June 2012. ISSN 1476-4687. doi: 10.1038/nature11234.
4. Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Ocean plankton. structure and function of the global ocean microbiome. *Science (New York, N.Y.)*, 348:1261359, May 2015. ISSN 1095-9203. doi: 10.1126/science.1261359.
5. Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, 103:12115–12120, August 2006. ISSN 0027-8424. doi: 10.1073/pnas.0605127103.
6. Sergio Balzano, Erwan Corre, Johan Decelle, Roberto Sierra, Patrick Wincker, Corinne Da Silva, Julie Poulain, Jan Pawlowski, and Fabrice Not. Transcriptome analyses to investigate symbiotic relationships between marine protists. *Frontiers in microbiology*, 6:98, 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00098.

7. Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, Jason W Sahl, Blaz Stres, Gerhard G Thallinger, David J Van Horn, and Carolyn F Weber. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*, 75: 7537–41, 2009. doi: 10.1128/AEM.01541-09.
8. J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7:335–336, May 2010. ISSN 1548-7105. doi: 10.1038/nmeth.f.303.
9. Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41:D590–D596, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1219.
10. Elizabeth M Glass, Jared Wilkening, Andreas Wilke, Dionysios Antonopoulos, and Folker Meyer. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor Protocols*, 2010(1):pdb–prot5368, 2010.
11. Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, and et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints*, 2018.
12. Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28:2520–2522, October 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts480.
13. Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3: e1420, 2015. ISSN 2167-8359. doi: 10.7717/peerj.1420.
14. Evguenia Kopylova, Laurent Noé, and Hélène Touzet. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics (Oxford, England)*, 28: 3211–3217, December 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts611.
15. Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, 26:2460–2461, October 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq461.
16. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22:1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.
17. Evguenia Kopylova, Jose A Navas-Molina, Céline Mercier, Zhenjiang Zech Xu, Frédéric Mahé, Yan He, Hong-Wei Zhou, Torbjørn Rognes, J Gregory Caporaso, and Rob Knight. Open-Source Sequence Clustering Methods Improve the State Of the Art. *mSystems*, 1, 2016. ISSN 2379-5077. doi: 10.1128/mSystems.00003-15.
18. S Andrews. Fastqc: a quality control tool for high throughput sequence data, 2010.
19. Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, 30:614–620, March 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt593.
20. Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, 17:10–12, May 2011. doi: 10.14806/ej.17.1.200.
21. Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016. ISSN 2167-8359. doi: 10.7717/peerj.2584.
22. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215:403–410, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.
23. Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73:5261–5267, August 2007. ISSN 0099-2240. doi: 10.1128/AEM.00062-07.
24. Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, December 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421.
25. J. Gregory Caporaso, Kyle Bittinger, Frederic D. Bushman, Todd Z. DeSantis, Gary L. Andersen, and Rob Knight. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics (Oxford, England)*, 26:266–267, January 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp636.
26. Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30: 772–780, April 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst010.
27. Eric P Nawrocki and Sean R Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)*, 29:2933–2935, November 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt509.
28. M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23:2947–2948, November 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm404.
29. Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32:1792–1797, 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh340.
30. Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22:2688–2690, November 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl446.
31. Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26:1641–1650, July 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp077.
32. Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12:385, September 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-385.
33. Laure Guillou, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, Gaétan Burgaud, Colomban de Vargas, Johan Decelle, Javier Del Campo, John R Dolan, Micah Dunthorn, Bente Edvardsen, Maria Holzmann, Wiebe H C F Kooistra, Enrique Lara, Noan Le Bescot, Ramiro Logares, Frédéric Mahé, Ramon Massana, Marina Montresor, Raphael Morard, Fabrice Not, Jan Pawlowski, Ian Probert, Anne-Laure Sauvadet, Raffaele Siano, Thorsten Stoeck, Daniel Vaulot, Pascal Zimmermann, and Richard Christen. The protist ribosomal reference database (pr2): a catalog of unicellular eukaryote small sub-unit rrna sequences with curated taxonomy. *Nucleic acids research*, 41:D597–D604, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1160.
34. J. D. L. van Bleijswijk, C. Whalen, G. C. A. Duineveld, M. S. S. Lavaleye, H. J. Witte, and F Mienis. Microbial assemblages on a cold-water coral mound at the SE Rockall Bank (NE Atlantic): interactions with hydrography and topography. *Biogeosciences*, 12:4483–4496, 2015.
35. Sergio Balzano, Julie Lattaud, Laura Villanueva, Sebastiaan W Rampen, Corina PD Brussaard, Judith van Bleijswijk, Nicole Bale, Jaap S Sinninghe Damsté, and Stefan Schouten. A quest for the biological sources of long chain alkyl diols in the western tropical North Atlantic Ocean. *Biogeosciences*, 15(19):5951–5968, 2018.
36. Marc A Besseling, Ellen C Hopmans, Michel Koenen, Marcel T.J. van der Meer, Sanne Vreugdenhil, Stefan Schouten, Jaap S. Sinninghe Damsté, and Laura Villanueva. Depth-related differences in archaeal populations impact the isoprenoid tetraether lipid composition of the Mediterranean Sea water column. *Organic Geochemistry*, 135:16–31, 2019.
37. Lise Klunder, Gerard CA Duineveld, Marc SS Lavaleye, Henk W van der Veer, Per J Palsbøll, and Judith DL van Bleijswijk. Diversity of Wadden Sea macrofauna and meiofauna communities highest in DNA from extractions preceded by cell lysis. *Journal of Sea Research*, 152: 101764, 2019.
38. Nature.com. Reporting standards and availability of data, materials, code and protocols. 2019.
39. Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS computational biology*, 9:e1003285, October 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003285.
40. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Cech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46:W537–W544, July 2018. ISSN 1362-4962. doi: 10.1093/nar/gky379.
41. LGJ van Bree, F Peterse, MTJ Van der Meer, JJ Middelburg, AMD Negash, W De Crop, Christine Cocquyt, JJ Wieringa, Dirk Verschuren, and JS Sinninghe Damsté. Seasonal variability in the abundance and stable carbon-isotopic composition of lipid biomarkers in suspended particulate matter from a stratified equatorial lake (Lake Chala, Kenya/Tanzania): Implications for the sedimentary record. *Quaternary Science Reviews*, 192:208–224, 2018.
42. Alma E Parada, David M Needham, and Jed A Fuhrman. Every base matters: assessing small subunit rrna primers for marine microbiomes with mock communities, time series and global field samples. *Environmental microbiology*, 18:1403–1414, May 2016. ISSN 1462-2920. doi: 10.1111/1462-2920.13023.
43. A. Apprill, S. McNally, R. Parsons, and L. Weber. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol*, 75(2):129–137, 2015.